SELF-GUIDED MOLECULAR DYNAMICS SIMULATION FOR EFFICIENT CONFORMATIONAL SEARCH

FIELD OF THE INVENTION

The present invention relates to Molecular Dynamics based simulation techniques. In particular, the invention relates to a simulation method capable of generating trajectories of enhanced conformational space sampling.

BACKGROUND OF THE INVENTION

Computer simulation techniques provide invaluable tools for the study of physical, chemical and biological processes. These techniques are based on microscopic descriptions of molecular systems which are processed to provide macroscopic information useful in understanding physical, chemical and biological phenomena. In this regard, processing molecular data generated by computer simulation provides unique opportunities for designing new molecular systems tailored to specific needs.

One popular computer simulation technique which employs microscopic description of a molecular system in predicting the system's macroscopic behavior is the Molecular dynamics (MD) technique. The MD simulation technique is a powerful computational tool which allows the exploration of structural, thermodynamic and kinetic properties of molecular systems. MD-based simulation methods have been successfully employed in the study systems varying from simple fluids to large proteins and DNA molecules.

Molecular dynamics provides a methodology for detailed microscopic modeling on the molecular scale of a molecular system containing N components (*N*-body). The conceptual basis for the physical relevance of data obtained through MD simulations is supported by the preposition that the nature of matter is to be found in the structure and motion of its constituent building blocks, and

therefore, the dynamic properties of a system are contained in the solution to the *N*-body problem. Given that the classical *N*-body problem lacks a general analytical solution, the only path open is the numerical one. Scientists engaged in studying matter at this level require computational tools to allow them to follow the movement of individual molecules and it is this need that the molecular dynamics approach aims to fulfill. This requires massive amounts of computing resources. In the last decade, the use of MD simulations in the study of physical, chemical and biological problems has increased exponentially, primarily due to the increased computing power of calculating machines.

The Molecular Dynamics simulation technique has become a popular approach to study many systems in physics, chemistry, and biology and predict their properties. Conformational sampling and searching lays the fundamental ground for MD simulations. The major limitation associated with conventional MD simulations is their extremely inefficient conformational searching and sampling efficiency, which severely limits the applicability of MD simulation and/or accuracy when predicting the properties of a system. Therefore, MD simulations have been primarily applied to homogenous systems or to fast events in heterogeneous systems. In many problems such as protein folding, ligand binding, molecular docking, phase transition, the conformation space of the system is so large and the time scale of an event is so long that insight that can be obtained by conventional MD simulation techniques is very limited even with today's fastest computer.

Molecular Dynamics techniques are based on the propagation of classical equations of motion. MD trajectories are generated by shifting the positions of the atoms forming the N-body system according to the forces exerted on each atom. At each new position, the interactions between a given atom and the remaining components of the molecular system are evaluated and derived to obtain the force

exerted on the atom at the given time step. That force is in turn employed in determining a new position for the atom. The position shifting and force evaluating is repeated along to form a trajectory. The trajectory is a collection of successive snapshots or frames separated in the temporal space by a time step of a given magnitude. The choice of a given time step is dictated by the nature of the system being simulated and is generally adjusted as a function of the complexity of the system as described, for example, by the complexity of the force field describing the interactions between the components of the simulated system and sometimes between those components and external elements. In general, the magnitude of a time step employed in a simulation is adjusted such that fluctuations in the energy of the simulated system between two successive time frames are small enough to allow the use of numerical approximations without drastically departing from purely theoretical trajectories. Acceptable time steps are generally a fraction of the time scale associated with significant changes in the topology of the energy surface from which the forces governing the motion being simulated are derived. Those changes are generally associated with the fastest movements. Therefore, the time step is generally adjusted to accurately describe the fastest movements in a molecular system.

Motions in molecular systems take place on a wide range of time scales. Vibrations of covalent bonds take place in 10^{-14} second, whereas large-scale conformational changes such as that in the protein folding process may take seconds or even hours. This large range of time scales presents MD simulations a great challenge. In an MD simulation, the classical equations of motion for all particles of a system are integrated over a certain period of time. Because of the limitation in the size of time steps used to solve the equations of motion, the time scale that an MD simulation can access is very short with current available computing power, typically in the range of pico-seconds (ps) to nano-seconds (ns).

This short time scale severely limits the applicability of MD simulations. To extend the applicability of MD simulations, a number of strategies have been used.

One field of intense research is the study of biological phenomena by molecular dynamics simulations. These systems present opportunities for the most important and spectacular utilization of molecular simulation in providing insights which can rarely be accessed through actual experimentation. In particular, possibilities of designing new molecules active as bio-agents through simulation is an avenue of great potential.

One of the biological problems that can be investigated by molecular dynamics simulation relates to employing the three dimensional structure of a protein in designing molecules that would bind specifically to the protein and thereby interfere with the activity of the protein. Since only a limited number of proteins having known tertiary structure are available, one important aspect in this regard relates to the prediction of a tertiary structure of a protein via computational methodology based solely on the amino acid sequence of said protein. Computer simulation techniques are important in this regard in view of the cost and time associated with the determination of such structure by experimental means such as NMR spectroscopy and x-ray crystallography which cannot be expected to keep pace with explosion of protein sequence information provided by DNA sequencing techniques. With the cost/performance of computational hardware being reduced by roughly a factor of two every eighteen months, and with improvements in software and algorithm development providing an even greater enhancement of computational efficiency, it is clear that the goal of computation-based structure determination will eventually be reached.

However, there are still formidable problems, both conceptual and numerical, in making this objective a reality. These problems involve both the issue of the appropriate representation of the protein (most critically, the potential

functions that assign energies to any given protein structure) and the algorithmic issue of finding the native structure, starting from an unfolded chain, in a finite amount of computation time. These two central areas are of course intertwined: if computational expense was irrelevant, one could simply solve the Schrodinger equation for each protein configuration (averaging over all positions of the solvent molecules), thereby guaranteeing accurate evaluation of the potential energy. The coupled problem is then to have the total computation time required to obtaining the predicted structure, roughly the cost of evaluating the energy of each structure multiplied by the number of such evaluations, be tractable with current computing technology. The development of new potential functions and the investigation of improved methods for the minimization of functions (to reduce the number of evaluations of the potential function that are necessary) are therefore two key areas in computational studies of protein folding.

The two most striking features of native protein structures in water are (a) the overwhelming tendency of charged or polar side chains to appear on the "exterior" of the protein (i.e. exposed to solvent) and hydrophobic side chains to be located in the "interior" of the protein (protected from solvent), forming a unique compact globular structure for a given sequence; (b) the formation of secondary structure segments, alpha-helices and beta-sheets, so as to satisfy backbone hydrogen bonds. In a naturally occurring globular protein, the pattern of hydrophobic and hydrophilic residues is such that the formation of a micellar structure, with the hydrophilic groups exposed to water and the hydrophobic groups buried in the interior, is possible. A central feature of proteins, as compared to other polymers or colloidal structures, is that this micellar formation must be compatible with the two secondary structure classes. If the backbone groups inside the protein cannot make a sufficient number of internal hydrogen bonds, the

protection of hydrophobic side chains from solvent will not compensate for the removal of the backbone carbonyl and amino moieties from aqueous solution.

Simple topological arguments would then suggest that a relatively large fraction of the protein (~40-70%) must be in either an alpha-helix or a beta-sheet. Furthermore, each individual secondary structure region cannot be too long (or else the protein would lack the ability to fold into a compact globular structure) and the "loop" regions joining the secondary structure elements also cannot be too long (or else the backbone residues in the interior of this region would be have difficulty efficiently forming hydrogen bonds, as suggested above). The picture that then emerges is one of rigid rod-like segments (ignoring for the moment the twists and deformations of alpha-helices and beta-strands) joined by flexible loops, with each segment of length between 3 and 30 residues. This description in fact characterizes the vast majority of proteins whose structures are available in the protein data bank.

It is well known from both theoretical and experimental results that secondary structure elements in proteins are marginally stable; indeed, helical and beta-sheet segments in proteins rarely retain their form when removed from the protein environment and studied as small peptides. Thus, protein folding is a highly cooperative phenomenon in which the final secondary structure pattern and tertiary architecture are determined by global optimization of the hydrophobic effect and backbone hydrogen bonding. The native structure is a result of a delicate balance of energetic terms; the overall thermodynamic stability of a native protein at room temperature is many orders of magnitude smaller than the total energy of the system. These features make the computational determination of protein structure extremely demanding, particularly with regard to the magnitude of the conformational searching problem.

The helix is a common secondary structural motif found in proteins, and the mechanism of helix-coil interconversion is key to understanding the protein-folding problem. One very direct approach to the mechanism of helix folding is through MD simulation. However, because of energy barriers and random walk, the helix folding process of peptides is a difficult task for the conventional MD simulation. MD simulation has so far been successful to study the relative fast process such as the folding dynamics of a pentapeptide in water and the conversion of 3_{10} helix to α-helix of a 16-residue, alanine-based peptide in water. For folding simulations of larger peptides, modified force fields had to be used to circumvent the energy barrier problem.

A simple strategy is to improve the computing power, and indeed, as computing power increases, more difficult problems can be studied. Another strategy is to improve MD simulation efficiency. For example, fast motions limit the size of time steps used to solve the equation of motion and make simulations very expensive. Therefore, when fast motions are not of major interest and are separable from other motions, constraints may be used to replace the degrees of freedom of fast motions, so that much larger time steps can be used. Other strategies to improve MD simulation efficiency include the use of multiple time steps to save computational load, the improvement of numerical integration methods and the use of solvation models to replace explicit solvent molecules. However, the improvement in MD simulation efficiency is far from sufficient to study slow events.

Another very productive strategy is to speed up a slow event so that the event can take place fast enough to be observed on a time scale accessible by an MD simulation. Simulated annealing method is a very successful example in this category. In this method, high temperature is used to increase the ability of the system to overcome energy barriers. The system is then slowly cooled down to

low temperatures and simulated at physiologically relevant conditions. A major advantage of the simulated annealing MD method is that the system can effectively overcome energy barriers and reach different conformational regions at high temperature. A drawback of this method is that it needs multiple simulations to find the global minimum state of the system and to search properly the conformational space of the system.

Therefore, there remains a great need for a computer simulation method which can provide a more complete exploration of the conformational space of a molecular system. Particularly, there remains a need for a method based on molecular dynamics which allows for generating trajectories which allow both a comprehensive sampling of the conformational space available to a molecular system while at the same time allowing the calculation of physically meaningful properties by processing the data generated in said simulations.

SUMMARY OF THE INVENTION

The present invention relates to a method of forming a generated conformation of a molecular system. The method comprises generating a self-guided molecular dynamics trajectory of the molecular system. The trajectory is initiated by assigning to each atom in the molecular system a set of initial parameters comprising an initial velocity vector and an initial conformation. The initial conformation is formed by disposing each atom in the molecular system is an initial position. The method further comprises propagating the molecular dynamics trajectory according to a computational procedure comprising for each atom i in the molecular system: (a) calculating a current force exerted on said atom i by deriving an energy function describing interactions between said atom i and other atoms in said molecular system;(b) determining a guiding force g_i describing an average along a portion of said trajectory of a force exerted on said atom i; (c) determining a current position and velocity of said atom i by adjusting a previous

position and velocity according to the equation $m_i a_i = f_i$, wherein m_i is the mass of atom i, a_i is an acceleration of atom i and f_i is a force obtained by adding said current force calculated in (a) and said guiding force determined in (b); (d) replacing said previous position and velocity of said atom i with said current position and velocity of said atom i;(e) repeating steps (a) to (d) for one or more iterations to generate said trajectory; (f) forming said generated conformation by disposing the atoms of said molecule in respective current positions obtained during a given iteration and (g) processing said generated conformation to determine a structural, energetic, thermodynamic or kinetic property of said molecular system.

One embodiment of the invention provides a method of determining said guiding force g_i comprising accumulating a force along a portion of said trajectory to obtain a cumulative force, dividing said cumulative force by the length of said portion of said trajectory to obtain an average cumulative force and multiplying said average cumulative force with a parameter λ to obtain said guiding force λg_i, wherein λ is not zero. The length of said averaging portion can be defined by a number of iterations. The guiding force parameter λ can be negative or positive. The parameters λ and t_L are adjusted according to the molecular system to be investigated by SGMD. For example, in SGMD simulations involving proteins, values of λ between 0.01 and 10.00 are expected to be suitable for conducting the SGMD simulations. Values of λ between 0.10 and 1.00 are most preferred.

In another embodiment of the invention, the cumulative force g_i is obtained by adding said current force exerted on said atom i as calculated in step (a) along said portion of said trajectory.

In yet another embodiment of the invention, the cumulative force is obtained by determining an external force exerted on said atom i obtained by deriving an energy function describing the interactions between said atom i and

other atoms in said molecular system which are not positioned within three covalent bonds or less from said atom i.

In a further embodiment of the invention, said cumulative force is obtained by determining for atom i and each atom j positioned within three covalent bonds from said atom i, an external force exerted on said atom i or atom j obtained by deriving an energy function describing the interactions between said atom i or atom j and other atoms in said molecular system which are not positioned within three covalent bonds or less from said atom i; for each atom j multiplying said external force with a mass ratio m_i/M_j to obtain a mass weighed external force; and adding said mass weighed external force obtained for each atom j to said external force exerted on said atom i to obtain said cumulative force.

In yet another embodiment, for each iteration propagating said molecular dynamics simulation, the current velocities obtained in step (c) are scaled in such a way that the total energy of said molecular system remains constant after the propagation.

In yet another embodiment, the invention provides a method of generating a tertiary structure of a protein of unknown tertiary structure having a known primary structure defined by a sequence of amino-acids forming said protein.

In another embodiment, the invention provides a method of docking a guest molecule into a host molecule comprising generating a self-guided molecular dynamics simulation, wherein the simulated molecular system comprises said guest and host molecules, and wherein the initial conformation describes said guest molecule positioned outside said host molecule and said molecular dynamics trajectory comprises a generated conformation wherein said guest molecule is positioned within said host molecule.

In yet another embodiment, the invention provides a method of generating a phase transition of a molecular system by self-guided molecular dynamics

simulation according to the invention, wherein said initial conformation describes said molecular system in an initial physical state and said generated conformation describes said system in a physical state different from said initial physical state.

In a further embodiment, the invention provides a method of generating a molecular dynamics trajectory of a molecular system comprising: generating a succession of snap shots, each snap shot describing a frame of said molecular system, and each snap shot is separated from a previous snap shot in said trajectory by a time step δt, said trajectory having a total duration of up to $t_{final} = N\delta t$, wherein N is a predetermined integer greater or equal to 1, and wherein generating said trajectory comprises (a) initiating said molecular dynamics trajectory by assigning to each atom in said molecular system a set of initial parameters comprising an initial velocity vector and an initial frame wherein each atom in said molecular system is disposed in an initial position; (b) propagating said molecular dynamics trajectory by a computational procedure comprising for each atom i by: (i) calculating a current force exerted on said atom i by deriving an energy function describing interactions between said atom i and other atoms in said molecular system; (ii) determining a guiding force g_i describing an average along a portion of said trajectory of a force exerted on said atom i; (iii) determining a current position and velocity of said atom i by adjusting said previous position and velocity according to the equation $m_i a_i = f_i$, wherein m_i is the mass of atom i, a_i is an acceleration of atom i and f_i is a force obtained by adding said current force calculated in (i) and said guiding force determined in (ii); (iv) replacing said previous position and velocity of said atom i with said current position and velocity of said atom i;(c) repeating step (b) for up to N-1 iterations to generate said trajectory and processing said trajectory to determine a structural, energetic, thermodynamic or kinetic property of said molecular system.

Yet another embodiment of the invention provides a method of generating a self-guided molecular dynamics treaectory, wherein a guiding force $g_i(t_p)$ at a snap shot $t_p = p\delta t$, wherein p is $\leq N$, is equal to a guiding force $g_i^{(s)}(t_p)$, wherein $g_i^{(s)}(t_p)$ is determined according to equation (1) as follows:

$$g_i^{(s)}(t_p) = (1 - 1/L)g_i^{(s)}(t_p - \delta t) + (1/L)m_i\sum_{j\epsilon Si}\{(f_j^{(s)}(t_p) + \lambda g_j^{(s)}(t_p - \delta t))/M_j\}$$

wherein $g_i^{(s)}(t_p)$ is the guiding force exerted on atom i at a snapshot t_p, $g_i^{(s)}(t_p - \delta t)$ is the guiding force exerted on atom i at a snapshot $(t_p - \delta t)$, L is a length of a portion of said trajectory employed in computing said cumulative guiding force, m_i, atomic mass of atoms i, respectively, S_i is a group formed by all atoms in said molecular system which are linked to atom i through three covalent bonds or less, M_j is the total mass of atoms in S_j, $f_j^{(s)}(t_p)$ is a force exerted on atom j by all atoms in said molecular system which are not in group S_i; λ is a predetermined guiding force multiplication factor.

In a further embodiment, the invention provides a method of generating a self-guided molecular dynamics trajectory, wherein determining a current position $r_i(t_p + \delta t)$ of an atom i at snap shot p +1, comprises adjusting a position of said atom i according to equation (2) as follows:

$$r(t_p + \delta t) = r(t_p) + v_i(t_p + \delta t/2)\delta t$$

wherein $r(t_p)$ is the position of atom i at a snap shot immediately preceding the snap shot $t_p + \delta t$ and $v_i(t_p + \delta t/2)$ is the velocity of atom i at half a time step δt separating snap shots p and p +1, wherein $v_i(t_p + \delta t/2)$ is calculated according to equation (3) as follows:

$$v_i(t_p + \delta t/2) = v_i(t_p - \delta t/2) + (f_i(t_p) + \lambda g_i^{(s)}(t_p))\delta t/m_i.$$

In yet a further embodiment, the invention provides a method of predicting an unknown folded structure of a polypeptide. The method comprises generating a self-guided molecular dynamics trajectory according to the invention, wherein said molecular system comprises said polypeptide and said initial frame is defined

by disposing said polypeptide in a random conformation; calculating a thermodynamic property along said trajectory; and extracting from said trajectory a selected frame associated with a predetermined criterion based on said thermodynamic property and forming a predicted conformation by assigning to each atom in said polypeptide a corresponding position in said selected frame. Said molecular system being preferably immersed in a box of water molecules.

In yet another embodiment, the invention provide a method for the refinement of experimental determined structures. With NMR determined constraints for x-ray determined low resolution conformation, this invention can generate high resolution structures.

BRIEF DESCRIPTION OF THE FIGURES

Figures 1 (a)-(c) show an all-atom model of the alanine dipeptide, bonded substructures of atom CT and bonded substructure of atom C_1, respectively;

Figure 2 shows ϕ-ψ dihedral angle distribution (contour map) of the alanine dipeptide obtained from a 10,000 ps MD simulation in vacuum at 700 K using the AMBER force field;

Figure 3 shows ϕ-ψ dihedral angle distribution (contour map) of the alanine dipeptide obtained from a 10,000 ps self-guided MD simulation in vacuum at 700 K using the AMBER force field. In the simulation, λ=0.1 and t_l=0.2 ps;

Figure 4 shows ϕ-ψ dihedral angle distribution of the alanine dipeptide obtained from a 5,000 ps MD simulation in explicit water at 300 K using the AMBER force field;

Figure 5 shows ϕ-ψ dihedral angle distribution of the alanine dipeptide obtained from a 5,000 ps self-guided MD simulation in explicit water at 300 K using the AMBER force field. In the simulation, λ=0.5 and t_l= 2 ps;

Figures 6(a) and 6(b) are time profiles of ϕ and ψ, respectively during MD and SGMD simulations of the alanine dipeptide obtained from 5,000 ps MD or

self-guided MD simulations in explicit water at 300 K using the AMBER force field;

Figure 7 shows snapshots of the conformations of the 16-residue peptide obtained from the 10,000 ps conventional MD simulation in vacuum at 300 K using the AMBER force field, starting from the fully extended conformation. For clarity, hydrogen atoms were not shown;

Figure 8 shows snapshots of the conformations of the 16-residue peptide obtained from a 10,000 ps self-guided MD simulation in vacuum at 300 K using the AMBER force field, starting from the fully extended conformation. For clarity, hydrogen atoms were not shown. In the simulation, λ=0.1 and t_l=0.2 ps;

Figure 9 shows snapshots of the conformations of the 16-residue peptide obtained from a 10,000 ps self-guided MD simulation in vacuum at 300 K using the AMBER force field, starting from a local energy minimum conformation obtained from the MD simulation in vacuum at 300 K, For clarity, hydrogen atoms were not shown. In the simulation, λ=0.1 and t_l=0.2 ps;

Figure 10 shows the potential energies of the 16-residue peptide during simulations in vacuum at 300 K using the AMBER force field. The points shown in this figure are sub-averages over a period of 100 ps. The dashed line is the result obtained from the conventional MD simulation; the dotted line is the result obtained from the self-guided MD simulation starting from a fully extended conformation; and the solid line is the result obtained from the self-guided MD simulation starting from "the trapped structure" obtained at the end of the 10,000 ps conventional MD simulation;

Figure 11 shows a system of the 16-residue peptide in 1879 explicit waters. Cubic periodic boundary condition was applied in the simulations;

Figure 12 contains snapshots of the conformations of the 16-residue synthetic peptide during a 200 ps conventional MD simulation in explicit water at 300 K using the CHARMM force field;

Figure 13 contains snapshots of the conformations of the 16-residue synthetic peptide during a 200 ps self-guided MD simulation in explicit water at 300 K using the CHARMM force field. For clarity, hydrogen atoms were not shown. In the simulation, λ=0.5 and t_l= 2 ps;

Figure 14 contains snapshots of the conformations of the 16-residue synthetic peptide during a 200 ps self-guided MD simulation in explicit water at 300 K using the AMBER force field. For clarity, hydrogen atoms were not shown. In the simulation, λ=0.5 and t_l= 2 ps;

Figure 15(a) shows the time profile of a gyration radius along a 10 ns conventional MD simulation of a 16-mer peptide;

Figure 15(b) shows the time profile of a gyration radius along a 10 ns SGMD simulation of a 16-mer peptide;

Figure 16 shows Ramachandran plot of the 16-mer peptide in 10 ns SGMD simulation;

Figure 17 shows the time profile of the helix content along a 10 ns SGMD simulation of a 16-mer peptide;

Figure 18 shows the total number of helical residues for each conformation in a 10 ns SGMD simulation of a 16-mer peptide;

Figure 19 shows population of 9 major conformational clusters;

Figure 20 shows the time profile of secondary structures along the first 400 ps of a 10 ns SGMD simulation of a 16-mer peptide;

Figure 21 shows a proposed helix forming initiation mechanism;

Figure 22 shows the time profile of a gyration radius and the time profile of the helix content along a 10 ns SGMD simulation of a 16-mer peptide performed with λ=0.3 and t_l=2 p;

Figure 23 shows the structure of YPGDV;

Figure 24 shows a ϕ-ψ dihedral angle distribution of YPGDV obtained from a 2 ns conventional MD simulation;

Figure 25 shows a ϕ-ψ dihedral angle distribution of YPGDV obtained from a 2 ns SGMD simulation;

Figure 26 shows the center conformation of nine major clusters observed during 2 ns SGMD simulations with λ=0.0 (conventional MD), 0.1, 0.2, 0.3,0.4 and 0.5;

Figure 27 shows conformational transitions among the nine clusters described in Figure 25;

Figure 28 shows the argon system with tetragonal periodic boundary conditions;

Figure 29 shows the time profile of the potential energy along a SGMD trajectory of 500 argon atoms;

Figure 30 is the initial configuration of the guest and host molecular system in a box of waters (For clarity, water molecules are not shown); and

Figure 31 is the predicted complex structure in comparison with the X-ray determined structure.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

The present invention provides a novel simulation method for generating trajectories which describe the evolution of a molecular system during a determined period of time. The method is based on the molecular dynamics technique and incorporates simulation parameters which significantly enhance the

conformational sampling capabilities of molecular-dynamics simulation technologies.

Specifically, the method of the invention includes calculating a guiding force and implementing such guiding force in the generation of a molecular dynamics trajectory. The guiding force is accumulated through the ongoing simulation. The guiding force enhances the systematic motions taking place during the simulation. The systematic motions are defined as the motions resulting form the average forces exerted on the simulated system along the molecular dynamics trajectory. That is, by averaging the forces exerted on the system during the simulation, short range random motions are cancelled out and the forces related to long range movements are enhanced.

The guiding force implemented in the method of the invention is obtained by averaging the classical force exerted on a given atom over a given portion of the generated trajectory. The averaged force is then employed in adjusting the position of the atom from one snap shot to the next. That is, at each time step, the position of the atom is adjusted according to a force which includes both the force generally employed in conventional molecular dynamics simulations and the guiding force, which is obtained by averaging the classical force over a portion of the trajectory being generated immediately before the current time step. Since the guiding force is calculated based on data accumulated during the simulation of the trajectory being generated, the method is referred to as Self-Guided Molecular Dynamics or SGMD. A detailed description of the basic principles and the computational technologies involved in its implementation can be found in "Self-Guided Molecular Dynamics Simulation for Efficient Conformational Search, Xiongwu Wu and Shaomeng Wang, *The Journal of Physical Chemistry B*, Vol. 102, Number 37, pages 7238-7250. The contents of the article and the references cited therein are herby incorporated by reference in their entirety.

The guiding force introduced into the molecular dynamics simulations according to the invention can be controlled by several parameters. Like the classical force employed in conventional MD and in SGMD, the guiding force is determined by the force field employed in generating the trajectory. In addition, the invention provides two main parameters for controlling the degree of enhancement of the systematic motions during an SGMD simulation. Those parameters are the length of the trajectory portion which is employed in computing the average force. The length of trajectory can either be express in terms of the number of time steps L included in calculating the average force exerted on an atom or the corresponding simulation time $L\delta t$, wherein δt denotes the time step employed in propagating the equations of motion of the molecular system.

The method of the invention provides important advantages. For example, a trajectory generated through SGMD provides a significantly enhanced search of the conformational space available to a molecular system while allowing providing physically significant structural, thermodynamics and kinetic data. The combination of enhanced conformational sampling and conservation of the physical relevance of the data obtained through SGMD simulations opens many of the applications that have been inaccessible to conventional MD techniques as well as some the available modified techniques.

For example, the SGMD technique can be advantageously employed in the investigation of protein folding and the prediction of the tertiary structure of a protein of known amino acid sequence and unknown structure. In biological systems, protein folding generally takes milliseconds to hours, far beyond the reach of conventional MD simulations with current computer power.

Homology modeling is a very popular and powerful means in predicting 3D structures of proteins of unknown tertiary structure based on the known 3D structure of a homologous protein, which is employed as the template. The

accuracy of this technique is intimately related to degree of homology between the amino acid sequences of the protein of unknown tertiary structure and the protein serving as a template. When the degree of homology is not sufficiently high, the refinement on the template is necessary to take into account the effect of the sequence mismatches on the structure of the protein. This requires refinement techniques which can comprehensively explore the conformational space available to the template, thereby enhancing the possibility of identifying the structure corresponding to the global energy minimum. As described in relation to the general protein folding applications of the SGMD method, the technique can be equally advantageously employed in homology based protein structure determination techniques by perform simulations to fold the regions that have very unreliable predictions from homology modeling e.g., regions with a very low degree of homology.

Another application which can be advantageously explored by the SGMD simulation technique of the invention relates to ligand binding and molecular docking. These aspects are subjects of great practical application in design, discovery and development of new therapeutic agents in pharmaceutical industry. The information how a drug binds and interacts with its target molecule such as an enzyme or a receptor is critical to the success of rational drug design. NMR and X-ray diffraction techniques are experimental means of providing such information but such methods are very expensive and painstaking. SGMD allows major improvements in the study of the host-guest dynamics. In particular, the method of the invention allows more flexibility to be introduced into the simulated molecular system by incorporating the degrees of freedom of both the guest (ligand) and host (receptor) thereby dramatically increasing the structural diversity of the possible host-guest associations. The advantages provided by SGMD in the investigation of guest-host dynamics and energetics will open the

doors to important application highly important fields, such as rational drug design.

Another advantage of the enhanced conformational sampling provided by the method of the invention related to the ability to describe complicated events, such as phase transition. Phase transition is an important process with fundamental importance in physics, chemistry and biology. Because phase transition such as crystallization of liquids is a process associated with significant free-energy barriers, conventional MD simulation has very limited application in the applied to simulate phase transition events, such as the crystallization process of simple liquids. In this regard, the SGMD simulation techniques advantageously provides a direct approach to investigate phase transition processes, and to predict phase states and structure of systems of interest.

Other avenues for advantageous implementation of the SGMD technique provided by the invention include multiple dimensional NMR, X-ray diffraction techniques and other molecular structure determination techniques based on the combination of simulation and spectroscopic data.

In the context of conformational searching, the motions of a system may be divided into random and systematic motions. While the random motions are important to the sampling of detailed conformational distributions of the system, the systematic motions of the system are more relevant to the conformational search of a large conformational space.

For a many-body system, the systematic motions are normally much slower than the random motions because of energy barriers and random walk. Energy barriers are created by non-uniform energy surfaces and random walk is due to the frequent collisions between atoms. One example of random walk is the diffusion of solute molecules in solution. The constant collisions of solute molecules with surrounding solvent molecules make the solute molecules go through a very

complicated and detoured path, which looks like a random walk. Consequently, the diffusion speed of a solute is therefore far slower than its actual thermal motion. The slow systematic motions are the bottleneck that limits the efficiency of a conformational search during an MD simulation. Therefore, if we can somehow specifically enhance the systematic motions, the efficiency of conformational searching may be improved.

The present invention relates to a new MD simulation method, termed the self-guided MD simulation method, with the primary goal of improving the conformational searching efficiency. In this method, an extra guiding force is introduced, in addition to the instantaneous force derived from the force field, into the equation of motion to speed up the systematic motion of the system.

THE SELF-GUIDED MD SIMULATION METHOD

Conformational searching is a basic approach in theoretical and computational chemistry and molecular dynamics [MD] simulation and is a powerful computational tool to study structural, thermodynamic and kinetic properties of molecular systems. This new MD simulation method was based upon a new equation of motion in which a guiding force was introduced, in addition to the instantaneous force, to accelerate the systematic motion. This guiding force may be calculated as a time average of the instantaneous force from the trajectory obtained from the very same MD simulation. Bonded substructures were defined for molecular systems in order to enhance effectively both the inter- and intra-molecular systematic motions.

The guiding effect can be adjusted through two parameters λ and t_l, a larger value of λ will speed up the systematic conformational changes and improve the overall conformational searching efficiency. However, a too large value of λ might result in more alteration in ensemble averages and the conformational distribution. A proper value of λ will significantly improve the overall

conformational searching efficiency and at the same time, will not significantly alter the thermodynamic properties of the system. Since t_l represents the averaging time, a relative small value should be chosen for systems with fast conformational changes and a relatively large value should be chosen for systems with slow conformational changes to achieve the best guiding effect.

Equation of motion with an enhanced systematic motion

The equation of motion used in the conventional MD simulation under the Newtonian law is shown in eq. (1).

$$m_i a_i = f_i \tag{1}$$

Wherein m_i is the mass of atom i, a_i is the acceleration of atom i and f_i is the force acting on atom i.

In a many-body system, most of motions change their direction frequently due to collisions and restraints, such as bond stretching and angle bending. The systematic motion of atom i may be described by its average motion during a certain time period.

Wherein v_i is the average velocity for atom i from time $t - t_l$ to time t, $r_i(t)$ is the coordinate of atom i at time t, $v_i(\tau)$ is the velocity of atom i at time τ.

According to eq. (1), the equation of the systematic motion can be expressed by eq. (3) as follows:

$$m_i \bar{a}_i(t) = \frac{m_i}{t_l} \int_{t-t_l}^{t} a_i(\tau) d\tau = \frac{1}{t_l} \int_{t-t_l}^{t} f_i d\tau = \bar{f}_i \tag{2}$$

Wherein $\bar{a}_i$ is the (average) acceleration of the systematic motion of atom i at time t, $a_i(\tau)$ is the acceleration of atom i at time τ and $\bar{f}_i$ is the average force.

Eq. (3) clearly shows that the acceleration of the systematic motion is governed by the average force.

Therefore, the introduction of the average force into the equation of motion as a guiding force can in principle alter the systematic motion of the system. In order to control the effect of introducing a guiding force into the equation of motion of simulated system, the invention provides a guiding factor λ which is introduced in eq. (3). With a guiding factor in eq. (3) and combining it with eq. (1), a new equation of motion can be obtained, as shown in eq. (4).

Combining the new equation (3) and equation (4) provides equation (4) which describes an enhanced systematic motion equation:

$$m_i A_i = m_i(a_i + \lambda \bar{a}_i) = f_i + \lambda \bar{f}_i \qquad \textbf{(3)}$$

Wherein A_i is the acceleration of atom i with altered systematic motion as compared to a_i in eq. (1), λ is a guiding factor, which can be any value. Since the acceleration of the systematic (average) motion is governed by the average force, a positive value of λ should speed up the systematic motion and a negative value of λ should slow down the systematic motion. A value of 0 for λ yield an eq. which is identical to the classical Newtonian equation of motion, described in eq. (1).

The implementation of eq. (4) with $\lambda>0$ provides an MD simulation with enhanced systematic motion. However, to perform an MD simulation precisely according to eq. (4), one needs to perform an MD simulation according to eq. (1) to obtain the average force. This needs to be done at every time step, which essentially makes an MD simulation according to eq. (4) impractical. To overcome this difficulty, eq. (4) was modified to obtain eq. (5) as follows:

$$m_i A_i = m_i a_i + \lambda m_i \bar{A}_i = f_i + \lambda g_i \qquad \textbf{(4)}$$

Where g_i is calculated from eq. (6) as follows:

$$g_i = m_i \overline{A}_i = \frac{1}{t_l} \int_{t-t_l}^{t} (f_i + \lambda g_i) d\tau \qquad (5)$$

While simulations generated according to equation (5) provide trajectories with enhanced systematic motion, the enhanced systematic motion includes inter- and intra-molecular systematic motions as discussed above, most of atomic motions related to intra-molecular interaction or bonded interactions are fast, while a systematic motion generally is very slow. During a conformational change, most of the internal coordinates, such as bond lengths, bond angles, and some of dihedral angles, just oscillate around their equilibrium values. Even though most of these high frequent motions do not contribute to a systematic conformational change, they create significant noise when using a time average velocity to describe the systematic motion. Hence, a guiding force estimated based upon the total force in a molecular system may be not very effective to accelerate the systematic motion. In one embodiment, the present invention provides bonded substructures to describe the local structural rigidity of a molecule and to derive a guiding force to enhance effectively the systematic motion.

For each atom, i, a bonded substructure, S_i is defined as a part of the molecule which includes atom i itself and all other atoms having a non-zero bonded interaction with atom i, i.e., atoms interacting with atom i, through either a bond, a bond angle or a dihedral angle. In other words, all atoms in a molecule to atom i through three or fewer covalent bonds belong to bonded substructure S_i. Atom i is called the central atom of substructure S_i. Each atom has a bonded substructure and two bonded substructures in a molecule can be partially or even totally overlapping as long as the central atoms are different. An example of substructures according to the invention is depicted in Figure 1. Figures 1(a)-(c)

depict an all-atom model for an alanine dipeptide and two bonded substructures for two atoms in the molecule.

Within the framework of the bonded substructures provided by the invention, a conformational change in a molecule can be viewed as the external motion of bonded substructures on a particular bonded substructure S_i and the internal motion within said bonded substructures S_i. The systematic conformational change of a molecule can be described by the external systematic motion of bonded substructures. Since the latter, the external motion is important for the systematic conformational change since selective enhancement of the external systematic motion of bonded substructures effectively improve the conformational searching efficiency for molecular systems in an MD simulation according to the invention.

To accelerate the external systematic motion of the bonded substructures, the invention provides a specific guiding force $g_i^{(s)}$, which applied to atom i. the specific guiding force is incorporated in the equation of motion of atom i according to equation (7) as follows: Analogue to eq. (5), the equation of motion for atom i with a guiding force $g_i^{(s)}$ can be written in equation (7) as follows:

$$m_i a_i = f_i + \lambda g_i^{(s)} \qquad \textbf{(6)}$$

It should be noted that a_i is different from that in the classical Newtonian equation of motion expressed in eq (1). As well, since the guiding force $g_i^{(s)}$ is used to accelerate the systematic motion of bonded substructures, $g_i^{(s)}$ can no longer be calculated according to eq. (6), the mathematical machinery associated with the motion of bonded substructures is described below.

For atom i, only the interaction force between atom i and atoms outside S_i is included in calculating the external force exerted on bonded substructure S_i.

This interaction force is denoted as $f_i^{(s)}$, which can be calculated according to eq. (8) as follows:

$$f_i^{(s)} = \sum_{j \notin S_i} f_{ij} \tag{7}$$

Wherein f_{ij} is the force exerted by atom i on between atoms i and j. The summation is over all the atoms that do not belong to bonded substructure S_i ($j \notin S_i$).

With an extra guiding force $\lambda g_i^{(s)}$, the contribution of the combined force on atom i to the acceleration of S_i is as follows:

$$M_i a_{S_i} = f_i^{(S)} + \lambda g_i^{(S)} \tag{8}$$

Where M_i is the total mass of S_i, a_{Si} is the acceleration of S_i contributed from the combined force on atom i. Such acceleration will apply to all the atoms in S_i. If we neglect the rotation of the bonded substructure, for atom j in S_i, the acceleration of atom j resulted from the above motion of S_i in eq. (9) is as follows:

$$a_{S_i} = \frac{1}{M_i}(f_i^{(S)} + \lambda g_i^{(S)}) \tag{9}$$

Because atom j may belong to multiple bonded substructures, the total acceleration of atom j, $a_j^{(S)}$ resulting from the external motion exerted on these bonded substructures is the summation of the eternal force over all the bonded substructures to which atoms j belongs. For each bonded substructure S_i to which atom j belongs, its central atom i should also belong to S_j ($j \in S_i$). Hence, we have:

$$a_j^{(S)} = \sum_{i \in S_j} a_{S_i} = \sum_{i \in S_j} \frac{1}{M_i}(f_i^{(S)} + \lambda g_i^{(S)}) \qquad \textbf{(10)}$$

Because $g_i^{(s)}$ is introduced to enhance specifically the systematic motion of atom i resulted from the external motion of bonded substructures, it can be thus calculated from eq. (12).

$$g_i^{(S)} = m_i \bar{a}_i^{(S)} = \sum_{j \in S_i} \frac{m_i}{M_j t_1} \int_{t-t_1}^{t} (f_j^{(S)} + \lambda g_j^{(S)}) d\tau = \sum_{j \in S_i} \frac{m_i}{M_j}(\bar{f}_j^{(S)} + \lambda \bar{g}_j^{(S)}) \qquad \textbf{(11)}$$

Wherein ($j \in S_i$) indicates that atom j belongs to substructure S_i, $a_i^{(S)}$ is the time average of the total acceleration of atom i associated with the systematic motion exerted on the bonded substructures S_i.

Hence, a self-guided MD simulation with an enhanced (altered) systematic motion of bonded substructures is performed according to eq. (7). Eq. (7) is called the equation of motion with an enhanced systematic motion for molecular systems. The force t_i on atom i is calculated exactly in the same manner as in a conventional MD. The guiding force on atom i is calculated according to eq. (12). It should be noted that for mono-atomic molecular systems, the bonded substructure becomes a single atom. In this case, eq. (12) reduces to eq. (6) and eq. (7) reduces to eq. (5).

Estimation of the guiding force

As discussed above, the estimation of a guiding force according to equation (6) can be time consuming and thus may offset the numerical advantages associated with the enhanced systematic motion procedure. Therefore, the invention provides a technique for minimizing the computational cost of determining the guiding force $g_i(t)$. Computational efficiency for the calculation

of the time average for any property P, can be achieved through implementation of eq. (13) as follows:

$$\bar{P}(t)=\frac{\delta t}{t_l}\sum_{\tau=t-t_l}^{t}P(t)=\frac{\delta t}{t_l}\left(\sum_{\tau=t-t_l}^{t-\delta t}P(\tau)+P(t)\right)=$$
$$\frac{t-\delta t}{t_l}\left(\frac{\delta t}{t-\delta t}\sum_{\tau=t-t_l}^{t-\delta t}P(t)\right)+\frac{\delta t}{t_l}P(t)\cong\left(1-\frac{\delta t}{t_l}\right)\bar{P}(t-\delta t)+\frac{\delta t}{t_l}P(t) \qquad \textbf{(12)}$$

Wherein P(t)is the time average of property P at time t, $P(\tau)$ is the value of P at time τ, δt is the size of one MD time step, and t_l is the averaging time.

Eq. (13) shows that the time average of P at the current time step, P(t) can be estimated from the time average of the previous time step, P(t - δt) and the value of P at the current step P(t). This approximation can effectively save both memory and cpu time.

Accordingly, the guiding force in eq. (12) can be approximately calculated according to eq. (14).

$$g_i^{(S)}(t)=\left(1-\frac{\delta t}{t_l}\right)g_i^{(S)}(t-\delta t)+\frac{\delta t}{t_l}\sum_{j\in S_i}\frac{m_i}{M_j}(f_j^{(S)}(t)+\lambda g_i^{(S)}(t)) \qquad \textbf{(13)}$$

t_l should be larger than or equal to the size of one MD time step, δt.

The self-guided MD simulation algorithm

A self-guided MD simulation is performed by numerically solving eq. (7). In our implementation, the leap-frog algorithm was employed, as shown in eq. (15) and (16) as follows:

$$r_i(t+\delta t)=r_i(t)+v_i\left(t+\frac{\delta t}{2}\right)\delta t \qquad \textbf{(14)}$$

$$v_i\left(t+\frac{\delta t}{2}\right)=v_i\left(t-\frac{\delta t}{2}\right)+(f_i(t)+\lambda g_i^{(S)}(t))\frac{\delta t}{m_i} \qquad (15)$$

Wherein $v_i(t)$ is the velocity of atom i at time t, $f_i(t)$ is the force on atom i at time t, $\lambda g_i^{(S)}(t)$ is the guiding force on atom i at time t and λ is the guiding factor.

Because of the extra guiding forces, now atoms move along a trajectory departing from the classical Newtonian dynamics. For micro-canonical ensemble simulation, the total energy of the system should be conserved. This can be accomplished by scaling the velocities of atoms in the system after each time step. The velocities at the next half time step, $t + \delta t/2$, are scaled according to eq. (17).

$$v'_i\left(t+\frac{\delta t}{2}\right)=\chi_E v_i\left(t+\frac{\delta t}{2}\right) \qquad (16)$$

Wherein χ_E is called the energy conservation scaling factor, which was introduced to make the total energy conserved during each time step.

Hence, with the scaling factor χ_E, the kinetic energy at $t + \delta t/2$ is:

$$E'_k\left(t+\frac{\delta t}{2}\right)=\frac{1}{2}m_i\dot{r}'^2_i\left(t+\frac{\delta t}{2}\right)=\chi_E^2 E_k\left(t+\frac{\delta t}{2}\right) \qquad (17)$$

According to eq. (7), the kinetic energy change at each time step without the scaling factor χ_E can be expressed as:

$$\delta E_k = E_k\left(t+\frac{\delta t}{2}\right) - E_k\left(t-\frac{\delta t}{2}\right) = \sum_i m v_i(t)\ \delta v_i = \sum_i m v_i(t) a_i(t)\delta t = \sum_i (f_i+\lambda g_i^{(S)}) v_i(t)\delta t \quad (18)$$

While the potential energy change is:

$$\delta E_p = -\sum_i f_i \delta r_i = -\sum_i f_i v_i(t)\delta t \quad (19)$$

With the scaling factor in eq. (17), the total energy is thus conserved. Hence, the kinetic energy change should be equal to the negative potential energy change.

$$\delta E'_k = E'_k\left(t+\frac{\delta t}{2}\right) - E_k\left(t-\frac{\delta t}{2}\right) = \chi_E^2 E_k\left(t+\frac{\delta t}{2}\right) - E_k\left(t-\frac{\delta t}{2}\right) = -\delta E_p = \sum_i f_i v_i(t)\ \delta t \quad (20)$$

From eq. (19) and eq. (21), the following equation can be derived to calculate the scaling factor χ_E:

$$\chi_E = \left(\frac{E_k\left(t-\frac{\delta t}{2}\right) + \sum_i f_i v_i(t)\delta t}{E_k\left(t-\frac{\delta t}{2}\right) + \sum_i (f_i+\lambda g_i^{(S)}) v_i(t)\delta t} \right)^{1/2} \quad (21)$$

Wherein $v_i(t)$ is calculated by

$$v_i(t) = v_i\left(t-\frac{\delta t}{2}\right)+\frac{\delta t}{2m_i}(f_i+\lambda g_i^{(S)}) \qquad (22)$$

For constant temperature simulation, Berendsen et al. proposed a velocity scaling method[21] that uses a velocity rescaling factor χ_B as follows:

$$\chi_B=\left(1+\frac{\delta t}{t_T}\left(\frac{T}{\mathrm{T}}-1\right)\right)^{1/2} \qquad (23)$$

to maintain a constant temperature T. Wherein T is the current kinetic temperature and t_T is a preset time constant. Here, the energy conservation scaling factor χ_E and Berendsen's velocity rescaling factor χ_B to obtain a velocity scaling factor χ_T, for constant temperature simulation based on the self-guided MD method of the invention as follows:

$$\chi_T = \chi_B\chi_E \qquad (24)$$

Alternative to the scaling method described above, a constant temperature simulation using the self-guided MD method can also be accomplished using the constraint method to keep the kinetic energy constant.[22-25] The constrained equation of the self-guided motion can be written as:

$$m_i a_i = f_i+\lambda g_i^{(S)} - \xi v_i \qquad (25)$$

Wherein ξ is a kind of "friction coefficient", which varies to constrain the kinetic energy to a constant value. This constraint method has also been implemented but was not used in the simulations presented in this paper. Thus, the detail algorithm of its implementation is not discussed here.

Based on the above mathematical framework for employing the "leap-frog" algorithms generating a self-guided MD simulation according to the invention, a computer simulation is performed as follows:

(a) Provide a set of initial conditions for starting the simulation. The initial conditions include for each atomists of simulated molecular system an initial position at t=0, an initial velocity v(t=0), a time step δt, a guiding factor λ and a guiding force averaging time, t_l and an initial guiding force $gi^{(j)}(0)$.

(b) At time step t, calculate the force on each atom, $f_i(t)$. This calculation is exactly the same as that in a conventional MD simulation. Calculate the guiding force according to (27) which is an approximation of eq. 14.

$$g_i^{(S)}(t)=\left(1-\frac{\delta t}{t_l}\right)g_i^{(S)}(t-\delta t)+\frac{\delta t}{t_l}m_i\sum_{j\in S_i}\frac{f_j^{(S)}(t)+\lambda g_j^{(S)}(t-\delta t)}{M_j} \qquad \textbf{(26)}$$

(c) Forward the velocities to the next half time step, according to eq. (15). For constant energy simulations, the velocities are scaled according to eq. (17). For constant temperature simulations, the velocities are scaled by the scaling factor calculated according to eq. (25).

(d) Forward the positions to the next time step, according to eq. (16). If bond lengths or other internal coordinates are to be fixed, the

SHAKE algorithm developed by Ryckaert et al.[6] is applied to get the constrained positions.

(e) Repeat steps (b) to (d) until the end of the simulation.

EXAMPLES

EXAMPLE 1

Simulation Conditions

In this example, the self-guided MD simulation method was applied to two peptides, an alanine dipeptide and a neutral, 16-residue synthetic peptide. The structure of the alanine dipeptide is shown in Figure 1(a). The alanine dipeptide system has been extensively studied by MD simulations and other computational methods. The neutral, 16-residue synthetic peptide has a sequence of CH_3CO-AAQAAAAQAAAAQAAY-NH_2, where A refers to an alanine, Q refers to a glutamine, and Y refers to a tyrosine. This synthetic peptide was determined experimentally using the circular dichroism (CD) spectrum to have ca. 50% helix secondary structure in aqueous solution. All-atom models were used to represent these peptides. For the purpose of comparison, we also carried out comparative simulations for these two peptides using the conventional MD method. Identical conditions were used when carrying out the conventional MD and the self-guided MD simulations except in the conventional MD simulations, the guiding factor λ was set to be 0.

The self-guided MD simulation algorithm was implemented in both the CHARMM and the AMBER programs. The MD simulations were performed using the CHARMM package with its version 22 force field parameters or the AMBER (version 4.0) with its all-atom force field parameters. Both sets of force field parameters include bond length, bond angle, dihedral angel and improper dihedral angle interactions for bonded interaction, Lennard-Jones 6-12 potential and electrostatic interaction for non-bonded interaction. The AMBER force field

also includes an explicit hydrogen bond potential. In all simulations, a constant dielectric constant of 1 was used. All results obtained using the AMBER force field are reported below. For brevity, the results obtained using the CHARMM force field are reported only for the simulation of the 16-residue peptide in explicit water at 300 K using the conventional MD and the self-guided MD simulation methods.

Fully extended conformations were used as the initial conformation for these peptides. The aqueous solution for a peptide was built by immersing the peptide into an equilibrated water box and deleting overlapping water molecules found within 2.45 Å of the peptide. The TIP3P water model[41] was used to represent water molecules.

For simulations conducted in vacuum i.e., with no explicit water molecules in the molecular system, non-bonded interactions were calculated with no cutoff. For simulations in explicit water, non-bonded interactions were calculated with a cutoff value of 14 Å and a switching function was applied from 8 to 12 Å to smooth the interaction change across the cutoff. Cubic periodic boundary conditions were applied to molecular systems including explicit water molecules. A time step δt of 0.002 ps was used in all simulations. A neighbor list was created for interaction calculation and updated every 10 MD time steps. The SHAKE algorithm was applied to constrain all bonds. All simulations were conducted at constant temperature. The velocity scaling method was used to maintain a constant temperature and t_T in eq. (24) was set to be 1 ps in all simulations.

(A) <u>The Dipeptide System</u>

1. <u>Alanine dipeptide simulations in vacuum</u>

The alanine dipeptide has a total of 22 atoms with two peptide bonds (Figure 1(a)) . Two dihedral angles, ϕ for C_1-N_1-C_α-C_2 and ψ for N_1-C_α-C_2-N_2, in this molecule can freely rotate and the rotation of these two torsion angles results in

different conformations for the molecule. The conformation of this molecule can be described by the values of ϕ and ψ. Conformations with different ϕ and ψ angles have different potential energy because of the interactions between atoms in the molecule. Therefore, the conformational distribution is not uniform throughout the ϕ-ψ space. The ϕ-ψ distribution is characteristic for each amino acid and this distribution can be shown by a Ramachandran plot, which depicts the conformational densities in a 2D plot with ϕ being the X-axis and ψ being the Y-axis.

Two comparative 10,000 ps simulations were performed to the peptide at 700 K with the AMBER force field. For the purpose of comparison, a temperature of 700 K was chosen for both simulations because it was found that at lower temperatures such as 300 K, a 10,000 ps simulation using the conventional MD method does not provide proper sampling of the entire conformational space and a much longer simulation time would be required to obtain a proper sampling. The parameters for the self-guided MD simulation were chosen as λ=0.1 and t_l=0.2 ps. Trajectories were recorded every 0.2 ps in both simulations for analysis.

First, the ensemble average properties obtained from these two simulations were evaluated. The ensemble averages of temperature, the fluctuation of temperature, the total energy, the fluctuation of total energy, the potential energy, the fluctuation of the potential energy, the kinetic energy and fluctuation of kinetic energy obtained from both simulations are provided in Table I. As can be seen, the values obtained from the conventional MD and the self-guided MD methods are essentially the same, indicating the self-guiding force does not change significantly these ensemble average properties.

Table I. Comparison of ensemble averages between the conventional MD and the self-guided MD methods for the alanine dipeptide systems.

	Temperature (K)		Total energy (kcal/mol)		Kinetic energy (kcal/mol)		Potential energy (kcal/mol)	
	Average	Fluctuation	Average	Fluctuation	Average	Fluctuation	Average	Fluctuation
Conventional MD (vacuum)	699.94±0.41	100.51±1.80	11.97±0.60	1.96±0.38	27.12±0.02	3.89±0.07	-15.16±0.59	4.54±0.17
Self-guided MD (vacuum)	699.69±0.46	100.73±1.89	11.86±0.69	1.84±0.41	27.11±0.02	3.90±0.07	-15.26±0.69	4.50±0.18
Conventional MD (water)	302.94±0.34	8.91±0.27	-1661.3±5.3	9.7±2.2	396.42±0.44	11.66±0.35	-2057.7±5.2	15.2±1.4
Self-guided MD (water)	303.15±0.34	8.88±0.28	-1658.2±5.1	9.7±2.1	396.69±0.45	11.63±0.36	-2054.9±5.1	15.1±1.4

Table II. Transition frequency between local minima in the conventional MD and the self-guided MD simulations.

	I↔II	I↔III	II↔IV	III↔IV
Conventional MD (vacuum)	10	302	59	29
Self-guided MD (vacuum)	31	374	139	35
Conventional MD (water)	11	29	3	1
Self-guided MD (water)	16	47	5	3

Next the conformational distributions in the ϕ-ψ space (the Ramachandran plot) obtained from conventional MD and the self-guided MD simulations are plotted in Figure 2 and Figure 3, respectively. Both maps reveal that there are four high-density regions, which are labeled as regions I to IV in Figures 2 and 3, respectively. These four high-density regions correspond to four local minima for the alanine dipeptide. In both contour maps, the ϕ and ψ values for these four local minima are (-75,75), (55, 35), (-65,-45), and (65,-75), respectively. According to the size of each region, region I was defined as the area within 50 degrees of peak I, region II as the area of 10 degrees within peak II, region III as the area of 10 degrees within peak III, and region IV as the area of 30 degrees within peak IV, respectively. No significant difference was found, when these two contour maps were superimposed. A small difference was observed in region II, where the self-guided MD simulation results in a sharper peak than the MD simulation. These results demonstrate that the conventional MD and the self-guided MD simulations yield almost identical results for the alanine dipeptide system in both the ensemble averages and the conformational distribution of the system.

The conformational searching efficiency of these two simulation methods was then compared. Four local minima were found in the conformational space as shown in Figures 2 and 3. Energy barriers exist between these four local minima, as evident from the low conformational density in the contour maps. The system needs to overcome these energy barriers in order to transfer from one region to another and to sample properly the ϕ-ψ conformational space during a simulation. Therefore, the frequency of the transitions between local minima is a measure of the ability to overcome the energy barriers as well as the conformational searching frequency. The transitions between local minima were counted for both simulations and the results were summarized in Table II. It is

clear that more transitions between the minima occurred in the self-guided MD simulation than in the conventional MD simulation. These results suggest that an improved conformational searching efficiency was achieved using the self-guided MD simulation method.

2. Alanine dipeptide simulations in aqueous solution

The alanine dipeptide was solvated using 212 TIP3P water molecules, enclosed in a cubic box. The box size was set to make the density of the system to be 1.0 g/cm^3. Two comparative 5,000 ps simulations were carried out at 300 K for the alanine dipeptide in aqueous solution using the conventional MD and the self-guided MD methods with the AMBER force field. Systematic motion in explicit water is generally much slower than that in vacuum and for this reason, a relatively bigger λ value of 0.5 was employed to increase the guiding effect of the self-guiding force and a relatively longer averaging time t_t of 2 ps was chosen for the self-guided MD simulation. Trajectories were recorded every 2 ps in both simulations for analysis.

The same ensemble average properties examined for the simulations in vacuum were evaluated for the solvated dipeptide system and the results are also provided in Table I. Again, the values obtained from the two simulations are essentially the same, indicating that the self-guiding force does not change significantly the ensemble average properties for the system in aqueous solution.

The conformational distributions in the ϕ-ψ space obtained from these two simulations are plotted in Figures 4 and 5, respectively. As can be seen in the figures, the conformational distribution obtained for these two simulations is similar. Region III, which is located at the same position, around (-60,-60) on both maps, has the highest conformational density. Two other regions with high conformational density, I and II, are also located at the same position on both maps. Region IV, which can be easily identified as a local minimum in Figure 5,

was only accessed a few times during the conventional MD simulation, as shown in Figure 4. The major difference between these two maps is that the conformations are more concentrated around the peaks in Figure 5 than in Figure 4 and this is particularly obvious for conformations in regions I and II. This suggests that the self-guided MD simulation accesses more often the low energy conformations than the conventional MD simulation. However, it is important to note that the positions of the global minimum and other local minima are the same in both maps.

The conformational searching efficiency for the simulations in aqueous solution was then compared. As can be seen from Figures 4 and 5, four regions of local minima were found in the conformational space and energy barriers that exist between these four local minima, as shown by the low conformational density in the contour maps. Four regions are defined as region I: 75 degrees around (-60,105); region II: 45 degrees around (45,60); region III: 50 degrees around (-60,-60); IV: 50 degrees around (60,-100). The transitions between the four regions were counted for both simulations and the results are also summarized in Table II. It is clear that more transitions between the four regions occurred in the self-guided MD simulation than in the conventional MD simulation.

Since each local minimum only covers certain range of ϕ and ψ angles, another way to examine the frequency of transitions between local minima is to examine the time profiles of ϕ and ψ, of which are plotted in Figures 6(a) for ϕ and 6(b) for ψ, respectively. As can be seen from Figure 6(a), more frequent transitions occurred for the ϕ angle during the self-guided MD simulation than during the conventional MD simulation and from Figure 6(b), similar results were obtained for the ψ angle. These results clearly showed that the guiding force in the

self-guided MD simulation method improves the conformational searching efficiency for the alanine dipeptide in the aqueous solution.

(B) The 16-Residue Synthetic Peptide System

When a system is large, the conformational space of the system becomes more complicated thus resulting during an MD simulation in the system being easily trapped in one local minimum or another. The proper sampling of the entire conformational space and the identification of the global minimum of a large system becomes an extremely difficult or even an impossible task to accomplish using the conventional MD simulation method at physically moderate temperatures. In this example, in order to illustrate the advantages provided by the self-guided molecular dynamics simulation for large systems, simulation studies on the synthetic 16-residue alanine peptide both in vacuum and in explicit water were conducted. This peptide is neutral, water soluble and has been determined to have ca. 50% helix secondary structure in aqueous solution.

1. Simulations in vacuum

Two comparative 10,000 ps simulations were performed at 300 K using the conventional MD and the self-guided MD simulation methods using the AMBER force field. In the self-guided MD simulation, λ of 0.1 and t_l of 0.2 ps were employed. Trajectories were recorded every 10 ps during the simulations for analysis.

In the conventional MD simulation starting from a fully extended conformation, the peptide collapsed to a compact structure within 100 ps (Figure 7). Continued simulation to 10000 ps did not result in much conformational change, indicating that the peptide was trapped in a local minimum. In the self-guided MD simulation at 300 K starting from the same fully extended conformation, the peptide folded into a helix structure within 100 ps (Figure 8). With continued simulation to 10,000 ps, significant conformational changes

occurred in local regions of the peptide but overall the peptide maintained the helix structure.

To further investigate the ability to overcome energy barriers, another self-guided MD simulation was performed starting from the "trapped structure" of the peptide obtained at the end of the 10,000 ps of conventional MD simulation. It was found that within 1,000 ps, the peptide folded into a complete helix structure (Figure 9). The helical structure obtained is similar to that obtained from the self-guided MD simulation starting from the fully extended conformation (Figure 7). Continued self-guided MD simulation to 10,000 ps showed that the peptide underwent some conformational changes in local regions but overall, the helix structure was maintained.

The potential energy of the peptide in these three simulations is plotted against the simulation time in Figure 10. As can be seen in the figure, in the conventional MD simulation, for the first 100 ps, the potential energy of the peptide decreased quickly and reached –505 kcal/mol. After 100 ps, the potential energy of the peptide fluctuated between –505 and –515 kcal/mol throughout the remaining period of the 10,000 ps simulation. In the self-guided MD simulation starting from the extended conformation, the potential energy of the peptide reached –520 kcal/mol and the complete helix formed within 100 ps. The potential energy of the peptide fluctuated between –520 and –530 kcal/mol throughout the remaining period of the 10,000 ps simulation. In the self-guided MD simulation starting from "the trapped structure" obtained from the conventional MD simulation, it was found that the potential energy of the peptide first went up to –495 kcal/mol from –506 kcal to overcome a large energy barrier then decreased quickly to –505 kcal/mol. At 500 ps, the potential energy went up again to –494 kcal/mol to overcome another large energy barrier and then decreased to –522 kcal/mol at ca. 1,000 ps simulation, when a complete helix was

formed. After this point, the potential energy of the peptide fluctuated between –520 and –530 kcal/mol, similar to the self-guided MD simulation starting from the fully extended conformation. From Figure 10, it can be seen that the potential energy jumps are much more frequent in the self-guided MD simulations than in the conventional MD simulation, which indicates more efficient conformational sampling by the self-guided molecular dynamics simulaiton.

2. Simulations in aqueous solution

Folding of the 16-mer peptide in aqueous solution through MD simulations would be difficult to achieve because the systematic conformational change of the peptide in aqueous solution is very slow due to the strong and random collisions of the peptide with its surrounding solvent molecules. Helix folding has been the subject of many simulation studies. As a result, de novo helix folding of large peptides in explicit water through MD simulations, is very hard to achieve, presumably because the systematic conformational change of large peptides in explicit water is very slow.

In illustrating the ability of the self-guided MD simulations according to the invention to describe conformational change in large molecular systems including explicit molecules, a total of 1879 TIP3P water molecules were used to solvate the 16-mer peptide in its extended conformation (Figure 11) and the density of the system was set to be 1.0 g/cm^3. The peptide was placed along the diagonal of the cube to minimize the box size. At the beginning of the simulation, the closest distance between the peptide and its images was 7.8 Å. As the simulation proceeded, the peptide began to fold and the closest distance between the peptide and its images became bigger. Overall, the interactions between the peptide and its images are not significant for the system.

Comparative, 200 ps simulations were performed at 300 K with both the CHARMM and AMBER force fields using both the conventional MD and the self-

guided MD methods. In the self-guided MD simulation, λ of 0.5 and t_1 of 2 ps were used. Each of these 200 ps simulations took ca. 80 CPU hours on the SGI indigo2 system with a single R10000 processor.

During the 200 ps conventional MD simulations using either the CHARMM or AMBER force fields, no significant conformational changes occurred for the peptide. Snapshots of the structure of the peptide during the 200 ps conventional MD simulation using the CHARMM force field are shown in Figure 12. The structure of the peptide was merely relaxed and its N-terminal end was bent slightly. To observe significant conformational changes of the peptide, much longer simulation time is required.

In contrast, significant changes in the conformation of the peptide were observed during the 200 ps self-guided MD simulation using either the CHARMM or AMBER force fields.

In order to illustrate the advantages of the self-guided MD technique, a detailed analysis of the conformational change of the peptide during the 200 ps self-guided MD simulation using the CHARMM force field is provided as follows. At 18 ps, the first hydrogen bond between the carbonyl group (CO) of A_9 (A_9 refers to Alanine residue at the 9th position in the sequence) and the amido group (NH) of A_{12} was formed. At 20 ps, another hydrogen bond between the CO of A_7 and the NH of A_{10} was formed and at 22 ps, the first α-helix turn was formed between residues 8 and 12. The formation of this α-helix turn promoted a hydrogen bond between the CO of A_{10} and the NH of Q_{13} at 24 ps. At 26 ps, a 3_{10} helix segment was formed from residues 11 to 14. This helix segment propagated toward the C-terminal and at 40 ps, the helix segment extended to residues from 8 to 16. This long helix segment began to break at 44 ps from the C-terminal and the helix propagated toward N-terminal at 50 ps. At 66 ps, a hydrogen bond was formed between the CO of A_1 and the NH of A_5. At 70 ps, the helix segment

shifted to residues from 5 to 11. At 74 ps, a 3_{10} helix was formed between residues 3 and 6 and another helix segment was formed from residues 6 to 12. At 80 ps, a long helix segment was formed from residues 1 to 13. At 86 ps, a helix was formed throughout the entire peptide with a slight distortion at its C-terminal. The helix structure was broken shortly after 100 ps into two helix segments, one from residues 1 to 7 and another one from residues 8 to 16. The conformational changes of the peptide continued to the end of the 200 ps simulation. A number of snapshots are shown in Figure 13. The results obtained from the 200 ps self-guided MD simulation using the CHARMM force field are consistent with the experimental results that showed that this peptide has ca. 50% helix structure in aqueous solution.

During the 200 ps self-guided MD simulation using the AMBER force field, the 16-residue peptide also folded into variety of helical segments, but the structural details differed from those obtained using the CHARMM force field. A number of snapshots of the structure of the peptide during the self-guided MD simulation using the AMBER force field are shown in Figure 14. The analysis of the conformations of the peptide during the simulation using the AMBER force field showed that the helix form is the dominant structural element of the peptide, again consistent with the experimental results.

Taken together, the comparative simulations on the 16-mer peptide in aqueous solution using the conventional MD and the self-guided MD methods clearly show that the self-guided MD simulation method provides a dramatic improvement in conformational searching efficiency. It is important to point out that the improvement in conformational searching efficiency does not depend upon the force field used in the simulation, although the structural details clearly depend upon the force field used in the simulation. Our results indicate that it is

feasible to study the folding dynamics of relatively large peptide systems in explicit water using the self-guided MD simulation method of the invention.

EXAMPLE 2

This example illustrates the advantages provided by the self-guided MD technique of the invention in analyzing folding mechanisms. One important step in protein folding processes is the formation of segments of secondary structure, such as α-helical segments. Therefore, analyzing the mechanisms associated with the formation of such secondary structures is an important step towards designing computational techniques capable of predicting protein territory structures.

Helix is a rudimentary secondary structure in proteins and helix formation has been found to play an important role in protein folding, which remains one of the most challenging and difficult problems in biological science. Elucidation of helix folding mechanism can therefore provide important insights into the protein-folding problem.

In this example, we performed a 10-ns SGMD simulation of a 16-residue, alanine-based peptide in explicit waters using a set of SGMD guiding parameters as described in Example 1. In addition, we performed a 10 ns simulation of the peptide under the same simulation condition with conventional MD method. Furthermore, to investigate the effects of the guiding parameters to the helix folding in the SGMD simulations, we performed another 10 ns SGMD simulation with different guiding parameters. It was found that both 10 ns SGMD simulations were able to achieve reversible helix folding and that the 10 ns conventional MD simulation was not sufficiently long for the formation of a single helix segment. Analysis of the SGMD simulation trajectories provides us the unprecedented opportunity to study the helix folding mechanism in water. The simulation conditions and results are discussed below.

Simulation Setup

In this example, the peptide aqueous solution was constructed by immersing the 16-mer peptide molecule in its fully extended conformation in a box of 1945 TIP3P waters and deleting overlapping water molecules that were within 2.45 Å of the peptide. Cubic periodic boundary conditions were applied in the simulation to eliminate the boundary effect and a SHAKE algorithm was used to fix all bond lengths during the simulation such that large time step of 0.002 ps can be employed. NTP simulations were performed (T=274K and P= 1 atm). Temperature and pressure are controlled using conventional algorithms as described above. The temperature was coupled to a temperature bath with a coupling constant of 1.0 ps, and the pressure coupling constant was set to be 1.0 ps. Conventional MD simulation was performed using the AMBER program, while SGMD simulations were performed using the SGMD method implemented in the AMBER program. AMBER94 all atom force field (Cornell et al. 1995) was used to represent the peptide. A constant dielectric constant of 1 was used for the electrostatic interaction calculations. The cut-off value for non-bonded interactions, including Lennard-Jones and electrostatic interaction was set at 10 Å based on residues.

Secondary Structure Analysis

To provide a detailed description of the secondary structure of the peptide during the simulation, the dictionary of secondary structure of protein (DSSP) proposed by Kabsch & Sander was employed to analyze the secondary structural elements of the peptide conformation. According to the DSSP definition, cooperative secondary structure is recognized as repeats of the elementary hydrogen-bonding patterns "turn" and "bridge". Repeating turns are "helices", repeating bridges are "ladders", connected ladders are "sheets". Geometric structure is defined in terms of the concepts of torsion and curvature of differential

geometry. Local chain "chirality" is the torsion handedness of four consecutive Ca positions and is positive for right-handed helices and negative for ideal twisted b-sheets. Curved pieces are defined as "bends". For the 16-residue alanine-based peptide, the two terminal blocking groups, the acetyl at the N-terminal and the amide at the C-terminal, are counted as separate amino acids when defined the secondary structure of residues 1 to 16.

A. Conventional MD Simulation

We have performed a 10-ns conventional MD simulation on the peptide system. Analysis of the simulation trajectory showed that no helix segment was formed during the entire 10 ns simulation. In Figure 15, the gyration radius was plotted against the simulation time. As can be seen, the lowest gyration radius reached during the simulation is around 9 Å. SGMD simulations showed that when substantial helix segments are formed within the peptide, the gyration radius decreases to a value between 6 and 8 Å. Thus, the conformational change of the peptide segments during the 10 ns conventional MD simulation is not large enough for the formation of helix segments and is certainly far from sufficient for reversible helix folding. Based upon the recent 1 ms simulation experiment performed in water by Duan and Kollman (1998), substantial helix segments are formed after 50 ns. It is thus expected that the helix formation of this alanine-based peptide in water might also occur in a similar time-scale. Using the conventional MD simulation method, simulation time in the order of hundreds of picoseconds may be required in describing reversible folding of the 16-mer peptide. Since each 10 MD ns simulation takes about 10 weeks of CPU time on an R10000 SGI workstation, it will require substantial computing resource to perform a few hundred ns simulations at the present time.

B. The First 10 ns SGMD Simulation

The first 10 ns SGMD simulation was performed with an averaging time t_l equal to 2ps and a guiding factor λ equal to 0.4.

The gyration radius of the peptide was also plotted against the simulation time (Figure 16). As can be seen, the gyration radius reached 8 Å within 100 ps. Thereafter, the gyration radius fluctuated between 6 and 9 Å. Based upon the gyration radius change, SGMD simulation with this set of guiding parameters appears to accelerate the conformational change of the peptide by more than 100-fold as compared to conventional MD simulation.

The Ramachandran plot obtained for the 16-mer peptide based upon the 5000 conformations recorded during the 10-ns SGMD simulation is shown in Figure 16. As can be seen, the peptide has essentially accessed all the allowed conformational regions observed in protein structures, indicating that during the 10-ns SGMD simulation, sufficiently large conformational space was sampled. The most dominant backbone conformation is around the ∝-helix region (-57, -47), followed by the 3_{10}-helix region (-50, -20). Interestingly, the left-handed ∝-helix region (57, 47) and the β-structure region (-140, 140) are also populated. A significant population is located in a large, random coiled conformational region (-30 to –100, 0 to 180).

To provide a detailed description of the "motion" of each residue of the 16-mer peptide during the simulation, we classified the 5000 conformations recorded during the simulation according (DSSP). The average ratio of each secondary structural element (∝-helix , 3_{10}-helix , π-helix , turn, bend, bridge and ladder) for each residue is calculated in the peptide based upon all the 5000 conformations recorded during the simulation and provided in Table III.

Table III. The ratios of secondary structural elements observed for each amino acid in the peptide.

Residue #	Bend	Bridge	Ladder	Turn	3_{10}-Helix	a-Helix	p-Helix	All Helix ($a+3_{10}+p$)
1	0	0	0	9.5	0.6	75.0	0.5	76.2
2	9.3	0	0	10.3	0.9	77.4	0.6	78.9
3	5.8	0.2	0.3	8.0	1.0	78.7	0.7	80.5
4	4.0	0	0.3	4.5	0.6	81.9	0.7	83.1
5	2.8	0.1	0	7.6	4.1	82.4	0.7	87.1
6	5.5	0	0	16.3	5.8	69.0	0.3	75.1
7	7.7	0.1	0.3	18.6	6.5	62.3	0.2	69.0
8	10.5	0.2	0.3	20.2	5.4	50.1	0.7	56.2
9	15.0	0	0	7.9	2.6	45.9	0.7	49.1
10	7.4	0	0	10.0	2.2	71.0	0.7	73.9
11	4.5	0	0	22.2	3.9	66.7	0.7	71.3
12	7.4	0	0	23.9	5.3	55.4	0.6	61.3
13	9.9	0	0	19.4	7.1	50.7	0.1	57.9
14	13.5	0	0	25.5	7.7	45.5	0.1	53.3
15	23.2	0	0	25.0	4.6	33.4	0.0	38.0
16	0	0	0	25.5	2.1	10.6	0.0	12.7
Avg.	7.9	0	0.1	15.9	3.8	59.8	0.5	64.0

As can be seen from Table III, on average, helix secondary structure, including ∝-, 3_{10}-, and π-helix, is the dominant form for the peptide and the average helix ratio is 64%. Although the peptide may have not sampled all possible configurations during the 10 ns SGMD simulation, it is noted that the average helix ratio obtained from the simulation is comparable to the experimental helix ratio of 50%, as estimated from CD experiments (Scholtz, et al., 1991). Among the three different helix elements, ∝-helix is the dominant form (59.8%) and there is a significant population of the 3_{10}-helix (3.8%). The least populated helix form for this peptide is π-helix (0.5%). Our results thus clearly show that the ∝-helix conformation is much more populated than the 3_{10}-helix conformation.

This is consistent with a previous study for this peptide system, which showed that ∝-helix is more stable than 3_{10}-helix in aqueous solution.

A number of recent experimental studies using double spin label electron spin resonance (ESR) have suggested that a significant population of the 3_{10}-helix conformation form may exist for short alanine-based peptides. Based upon our simulation studies, it is clear that there is indeed a significant population of the 3_{10}-helix conformation (3.8%) for this peptide in solvated form. Thus, the UK simulation provides direct evidence for the existence of the 3_{10}-helix for alanine-based peptides in water. It is of note that the 3_{10}-helix conformation was mainly observed for two segments, residues 5 to 8 in the middle of the peptide and residues 11 to 15 near the C-terminus. Using a double spin label ESR technique, Fiori *et al.* have found that there is a strong evidence of 3_{10}-helix geometry near the C-terminus for 16- and 21-mer alanine-based peptides and our results are in good agreement with those experimental observations.

"Turn" is the next most populated secondary structural element. This structural element is believed to play an important role in helix initiation and propagation. Indeed, the self-guided MD simulation conducted according to the invention show that 87.5% of helix transitions (folding and unfolding) are between helices and turns, confirming the important role played by turns in helix folding. We observed a small population of bridges and β-strands, which mainly involve residues 3, 4, 7, and 8. Their low populations suggested that they are transient states under the simulation conditions. Left-handed helix was also observed from 6130 ps to 6560 ps, located at the C-terminus.

It is of interest to note that the helix ratio for each residue depends upon not only the nature of the residues (Ala, Tyr, Gln) but also the location of the residues. Very low helix ratios were observed for the C-terminal residues 15 and 16, being only 38.0% and 12.7%, respectively. From the N- to the C-terminal, two peaks of

the helix ratio were found, located at residues 5 and 10 and separated by residue 9. To further investigate the helix folding of the peptide during the simulation, for each conformation, each residue was classified as helical or non-helical according to the DSSP. The helix content of the peptide for each conformation was plotted against the simulation time (Figure 17). The helix segments were formed within the first 100 ps of the simulation. During the 10 ns SGMD simulation, helix folding, unfolding and refolding occur frequently, suggesting that the peptide structure was not trapped in a particular conformation. Figure 18, shows the total number of helical residues for each conformation against simulation time. As can be seen, within the first 500 ps, the peptide formed almost a complete helix, except for one residue break in the middle of the peptide. The peptide then partially unfolded around 1000 ps and refolded thereafter. At 2400 ps, the peptide completely unfolded. It refolded again into almost a complete helix with a frayed C-terminus around 3000 ps. At 4000 ps, the peptide completely unfolded again but refolded into a complete helix with a frayed C-terminus at 5200 ps. During the 10-ns SGMD simulation, the peptide completely unfolded 4 times. These data clearly showed that the reversible folding (folding, unfolding and refolding) of the peptide is achieved during the 10 ns SGMD simulation.

Based upon our simulation, it is clear that the peptide doesn't simply adopt either a complete helix or random coil conformations but has much rich conformational forms. We analyzed the major conformational clusters according to helix (H) and coil (C) definition. A total of 9 major conformational clusters were identified, as shown in Figure 19. As can be seen, the complete helix conformation only accounts for only 0.1% of the total number of the conformations. The coiled conformation (C) also accounts for 0.8% of the conformations. The most populated conformational cluster (HC, 38%) has a helix segment at the N-terminal and a frayed C-terminal. The next two most populated

conformational clusters have either one (CHC) or two helix segments (CHCHC) in the middle of the peptide and with frayed N- and C-terminals, accounting for 8.9% and 7.9% of the total conformations, respectively. The next most populated conformation (HCH) has two helix segments at the N- and C-terminal and with a break in the middle of the peptide (5%). The conformation with a C-terminal helix segment and a frayed N-terminal (CH) accounts for 4% of the total conformations. The conformation with a C-terminal helix segment, followed by a coiled segment, a helix segment near the N-terminal and a frayed N-terminal (CHCH) accounts for about 2% of the total population. Lastly, the conformation with a helix segment at the N-terminal, followed by a coiled segment, a small helix segment and a frayed C-terminal (HCHC), accounts for less than 1% of the total conformations. Our results thus showed that the alanine-based peptide adopts multiple conformations and the complete helix conformation only accounts for a very small percentage of total conformations. Accordingly, the experimentally observed 50% helix content for this alanine-based peptide should represent an average of several major conformational forms, each of which has one or more helix segments.

The self-guided MD trajectory was further analyzed as follows. Helix initiation starting from the fully extended conformation was first examined. Figure 20 shows the secondary structure evolvement during the first 400 ps of the simulation. As can be seen in the figure, starting from the fully extended conformation, the peptide first forms bends and turns. Thereafter, some turn structures further evolved into 3_{10}-helix segments. Segments of 3_{10}-helix can either transfer back to turns or further evolve and form α-helices.

In order to show that more precise understanding of the helix initiation mechanism can be gained through self-guided MD simulation, the total number of transitions between different secondary structural elements were analyzed. During the 10 ns simulation, there were a total of 250 transitions between 3_{10}-helix

segments and turns, 150 transitions between ∝-helix segments and turns, 17 transitions between π-helix segments and turns and 120 transitions between 3_{10}-helix and ∝-helix. There were only 9 transitions between bends and 3_{10}-helix segments. These results thus show that turns play an important role in helix initiation. Furthermore, our results indicate that there are more frequent transitions between turns and 3_{10}-helices than between turns and ∝-helices, although ∝-helix is a much more populated secondary structural element than 3_{10}-helix for this peptide during the simulation. However, there are also frequent and direct transitions between ∝-helix and turns. Therefore, starting from turn conformations, ∝-helix can be formed directly or through the 3_{10} helix. Based upon the foregoing results, a general helix initiation mechanism is proposed, as shown in Figure 21.

Previously, Millhauser proposed a helix folding mechanism based upon experimental data obtained from X-ray crystallography, CD, NMR and ESR and theoretical and simulations studies (Millhauser, 1995). The proposed key initiation mechanism show in Figure 21 is in excellent agreement with a mechanism based on X-ray crystallograph, CD, NMR and ESR experimental data as proposed by Millhauser. It is of note that the nascent helix conformation in Millhauser's helix folding mechanism resembles closely the turn conformation identified through the self-guided MD simulation of the invention. In both models, nascent helices or turns were formed first from random coiled conformations. The nascent helix further evolves into 3_{10}-helix, which can subsequently form ∝-helix. Both models identified the important roles played the nascent helix (turn) and the 3_{10}-helix in the folding of ∝-helix. One important extension in our helix folding model to Millhauser's helix folding model is that although turns (nascent helices) more readily fold into 3_{10}-helices, a significant percentage of turns conformations can fold into ∝-helices without going through the 3_{10}-helice as the intermediate.

C. **The Second 10-ns SGMD Simulation with Different Guiding**

Parameters

The impact of the guided motion parameters on the results of a self-guided MD simulation were investigated as follows:

A 10-ns SGMD simulation of the 16-mer peptide based on a different set of guiding parameters (an averaging time t_l=2 ps and a guiding factor λ=0.3) was carried.

In Figure 22, the gyration radius and the total number of helical residues are plotted against the simulation time. Based upon the rate of the gyration radius change, it is clear that the conformational change of the peptide during this SGMD simulation with a smaller guiding factor λ of 0.3 is significantly slower than that during the SGMD simulation with a guiding λ factor of 0.4, but much faster than the conformational change of the conventional MD simulation. Based upon the total number of helical residues, the peptide formed substantially α-helix at 3500 ps and reached almost a complete helix at 4000 ps, similar to the conformation observed at 430 ps of the SGMD simulation with a guiding factor λ of 0.4. During the 10 ns SGMD simulation with a guiding factor λ of 0.3, folding, partial unfolding and refolding were observed. However, the peptide didn't experience complete unfolding, unlike that observed during the 10 ns of the SGMD simulation with a guiding factor of 0.4. The helix profile of the 10-ns SGMD simulation with a guiding factor λ of 0.3 is very similar to that of the first 2-ns SGMD simulation with a guiding factor λ of 0.4. Nevertheless, both SGMD simulations showed that the peptide conformation is predominantly helix in explicit water, consistent with experimental results.

During the 10-ns SGMD simulation with a guiding factor λ of 0.3, helix folding, partial unfolding and refolding occurred many times and this provided another opportunity to examine the helix initiation mechanism and to validate the

results obtained from the 10-ns SGMD simulation with a guiding factor λ of 0.4. It was found that again turns play an important role to initiate helices (∝-helix, 3_{10}-helix and π-helix). During the 10-ns SGMD simulations, there were a total of 176 transitions between turns and 3_{10}-helices, 112 transitions between turns and ∝-helices and 36 transitions between turns and π-helices. For transitions involving ∝-helix, there were 112 transitions between turns and ∝-helices, 41 transitions between 3_{10}-helices and ∝-helices, and 4 transitions between π-helices and ∝-helices. These results show that a-helices are mainly initiated through the formation of turns. However, turns are more easily transferred to 3_{10}-helix segments and 3_{10}-helices also played an important role in the folding and unfolding of ∝-helices.

Comparison of the two SGMD simulations with different guiding parameters showed that with respect to the helix initiation, although the absolute number of transitions between turns, ∝-helices, 3_{10}-helices and π-helices differ between these two simulations, both simulations clearly indicate that turns play a dominant role in the formation of helices (∝-, 3_{10}- and π-helices). Moreover, both simulations clearly show that turns most likely transfer to 3_{10}-helices, while least likely transferring to π-helices among different forms of helices. While most ∝-helices are directly formed from turns, a significant percentage of ∝-helices are initiated from 3_{10}-helices. Furthermore, both simulations indicated that helix is the dominant conformational form of this alanine-based peptide in aqueous solution, consistent with experimental results.

EXAMPLE 3

In order to further examine the impact of the guiding parameters on a self-guided molecular dynamics simulation, the following computer experiments were performed on a system including the peptide YPGDV. The chemical structure of YPGDV is shown in Figure 23. The peptide was solvated using 805 TIP3P water

molecules in a cubic box and the box size was set to be 29Å. Cubic periodic boundary conditions were applied in all simulations. The CHARMM22 force field was employed. The nonbonded interactions were smoothly truncated. The cutoff distance for non-bonded interactions was set at 13 Å and a switch function was applied between 8 to 12 Å to smooth the change across the cutoff. A neighbor list was built and updated every 10 MD steps for non-bonded interaction calculation. The simulation temperature was set at 300 K. The SHAKE algorithm was employed to constrain all bonds so that a time step of 0.002 ps can be utilized in the simulations. The trajectories were recorded every 2 ps for later analysis. For the SGMD simulations, the local averaging time t_l is set at be 2 ps. To examine the guiding effect on conformational search, five SGMD simulations were performed with the guiding factor λ=0.1, 0.2, 0.3, 0.4, and 0.5, respectively.

1. Conformational space searched during 2 ns MD and SGMD simulations

In a 2 ns MD simulation starting from an extended conformation, the peptide experienced very slow conformational change. The water molecules surrounding the peptide formed a hydration shell presenting large energy barriers that prevent certain conformational transition and exerts damping effects on the movements of the peptide. A Ramachandran plot of ϕ and Ψ dihedral angles around each peptide bond separately is shown in Figure 23. As can be seen, of the eight dihedral angles, only ϕ_3 and ϕ_4 angles reached their whole range, while other angles just a oscillated around their starting region. Figure 24 shows some typical conformations obtained during the simulation. Starting from extended conformation (0 ps), the peptide relaxed to coil conformations such as the one at 370 ps with bends on it. It evolved into conformations with a type I turn like strucure over residues 2 to 5. After about half a nanosecond, this type I turn like

structure dispeared and the peptide remained in a coiled conformation for the rest of the simulation.

In contrast, in a 2 ns SGMD simulation with t_i=2 ps and λ=0.5, much wide regions have been searched (Figure 25). In addition to the regions reached in the MD simulation, f_3 found its high population region around 90°. ϕ_4 and ϕ_5 reached the region (0,90). Figure 26 shows some typical conformations obtained in the simulation. From the same extended conformation, the peptide quickly bended at residues 2 and 3 (at 32 ps) and formed a type II turn structure on residues 1-4 (34 ps). This type II turn structure is a fairly stable one and remained for the most of the simulation period. The conformational change was then localized in the C-terminus, which reached out to interact with water (120 ps) or folded back to interact with the N-terminus region (304 ps).

2. Acceleration of Conformational search in SGMD simulations

A series of 2 ns SGMD simulations with different guiding factors, λ=0.0, 0.1, 0.2, 0.3, 0.4, and 0.5. Here, λ=0.0 corresponding to the MD simulation. With these different guiding factor simulations, we can see the conformational space explored increases to a limit, and the conformational transition reach a reversible process.

To describe the conformational space accessed in these simulations the conformational clustering technique presented above is employed. Figure 26 shows the center conformation of nine major clusters observed in these simulations. The backbone ϕ, ψ dihedral angles are listed in Table IV. The conformational transition among clusters during these simulations are shown in Figure 27.

Table IV. Potential energies (kcal/mol) of the coil state and folded state in MD simulations.

Status	Et	Eu	Euv	Ev	Easp
Coiled	-8691.8±0.9	-75.7±0.3	-575.2±1.2	-8645.7±1.1	-29.7±0.1
Folded	-8691.7±0.9	-246.0±0.3	-246.3±0.6	-8461.3±0.9	-15.5±0.1

Easp: Solvation energy based on solvent accessible area using Honig et al.'s parameter.
Et: Total potential energy of the system.
Eu: intra-peptide potential energy.
Euv: Peptide-water interaction.
Ev: Water-water interaction.

Table V. Dihedral angles of the cluster central conformations from the SGMD and MD simulations and the dihedral angles for the ideal type I & II turn.

Clusters	y_1	f_2	y_2	f_3	y_3	f_4	y_4	f_5
1	160.03	-49.32	118.18	133.61	-65.85	-84.37	-128.78	-111.46
2	172.90	-60.86	117.47	82.43	-89.60	-57.61	156.23	-56.99
3	136.91	-38.79	123.00	90.71	-86.93	-77.72	85.02	38.51
4	152.14	-59.67	128.62	67.36	57.44	-69.78	121.31	-141.80
5	149.29	-58.47	-1.26	-65.67	-46.71	-88.63	-103.03	-137.09
6	148.91	-44.34	123.55	-68.32	147.40	-69.92	131.93	-140.08

7	146.65	-49.69	125.77	-81.77	-109.18	-61.29	-50.48	-122.94
8	176.41	-67.98	-3.64	-82.31	-150.22	-66.76	-54.44	-101.09
9	134.30	-53.23	130.68	-31.73	99.89	18.57	69.07	-100.79
Ideal type I turn on 2-5	¾	¾	¾	-60	-30	-90	0	¾
Ideal type II turn on 1-4	¾	-60	120	90	0	¾	¾	¾

Table VI. NOE data obtained from the NMR experiment and estimated from the SGMD simulation.

Atom pairs*	Residue pairs	NOE Intensity	
		NMR Experiment	SGMD Simulation
aN(*i*,*i*+1)	2-3	Very strong	8.54 (very strong)
	3-4	Moderately strong	2.24 (strong)
	4-5	Strong	3.38 (strong)
aN(*i*,*i*+2)	2-4	Weak	0.20 (medium)
NN(*i*,*i*+1)	3-4	Medium	0.88 (medium)

*aN(*i*,*j*) represents the atom pair of Ca proton and the amide proton in residues *i* and *j*, and NN(*i*,*j*), represents the atom pair of the amide protons in residues *i* and *j*.

did not result in reversible conformational change. That is, the peptide did not have repeated and stable visits to a cluster. At λ=0.4, repeat and stable conformational transitions between cluster II and III were observed. When λ=0.5, the peptide transferred to cluster IV after 32 ps and stayed there for about 100 ps before transferring to cluster I.

The above results show that for SGMD simulations to provide enhanced search (λ=0.3, 0.4, and 0.5) should be employed.

3. Comparison with NMR experimental results

The solution conformation of the pentapeptide, YPGDV, has been extensively studied using rotating frame NOESY(ROESY) NMR techniques. Their study showed that the Asp4 amide proton temperature coefficient for this peptide was only 3.3 ppb/K, considerably lower than that for other pentapeptides investigated in their study. This suggested that a high population of β-turn in solution might exist. Indeed, the NOE that was observed for this peptide confirmed its β-turn conformation in solution. For example, a strong NOE between the Gly_3 and Asp_4 amide protons was observed. The sequential cross peak between the NH of Asp_4 and the C∝H of Gly_3 is reduced in intensity as compared to that between the Asp_4 CaH and Val_5 NH and between Pro_2 CaH and Gly_3. Of great importance, a cross peak between the CaH resonance of Pro_2 and the NH resonance of Asp_4 was observed. Taken together, the very low Asp4 NH temperature coefficient and the NN(3,4) and ∝N(2,4) NOEs provide unequivocal evidence for a relatively high population of a 1-4 hydrogen bonded β-turn in this peptide. Formation of β-turn by this peptide was also confirmed independently by circular dichroism. The presence of a significant population in water solutions of this peptide is indicated by the positive ellipticity with a maximum near 206 nm and a minimum near 190 nm..

In the MD simulation (λ=0.0) starting from an extended conformation which belong to the coil cluster (VI), the peptide transferred to cluster VII, which has a type I-like turn structure on residues 2 to 5, after 150 ps and remained there for about 500 ps. After 630 ps, the peptide briefly visited cluster I region for a couple of times and back to cluster VI and remained there for the rest of period with occasional visits to cluster VII. In the MD simulation, starting from a type II turn structure (cluster I), the peptide remained in cluster I all the time except in two cases where it became close to cluster VII. The lack of transitions in this simulation indicates than cluster I is likely a stable state for this peptide.

When λ=0.1, the acceleration is not sufficient to make the peptide reach other important clusters in the 2 ns simulation. The peptide only briefly visited cluster VII and other clusters. A brief visit means that the peptide takes a conformation more similar to that in the cluster, but lack of stabilizing factors, such as hydrogen bonds and hydrophobic packing results in the peptide moving away from said cluster. At about 700 ps, the peptide transferred to cluster V, in which the terminus interaction induces two turn structures, one between residues 1-4 and the other one between residues 2-5. Residues 2, 3, and 4 formed a 3_{10} helix segment. In the rest simulation, the peptide remained in this cluster. When λ=0.3, the peptide reached cluster IX after 30 ps of simulation on time and remained there for more than 100 ps before transferring to cluster IV.

When λ=0.4 the peptide transferred to cluster X and stayed there for about 100 ps, then changed to cluster I at about 200 ps. In the next 600 ps, it stayed in cluster I with some brief visits to other clusters. After 900 ps the peptide transferred to cluster II and then to cluster III, and transfered between the two clusters during the rest of the period. From Table V we can see clusters I-IV all have a type II reverse turn on residues 1 to 4. But their C-termini have different positions relative to the turn structure. In simulations with λ<0.5, 2 ns simulation

First, the central conformations obtained by SGMD were compared with the NMR experimental results. As can be seen from Figure 26, clusters I-IV have a strong 1-4 hydrogen to bond between the carbonyl group of Tyr_1 and the amide group of Asp_4, in good agreement with the low temperature coefficient of the amide protein of Asp_4 and the NOE between 2H∝ and 4HN measured in the NMR experiment.

In NMR experiments, NOE represents average structural information in solution, rather than a specific conformation. To provide a more vigorous validation, the average NOEs were estimated from the SGMD conformations.

Using the 1000 conformations recorded during the 2 ns SGMD simulation with λ=0.5, the ensemble average NOE effects between each relevant proton pair are calculated and listed in Table VI. As can be seen, the agreement between SGMD simulation results and the experimental NOEs is remarkable.

EXAMPLE 4

This example illustrates the simulation of a phase transition by self-guided molecular dynamics according to the invention.

The SGMD simulation method was applied to L-J argon systems to examine its behavior and to demonstrate its ability in conformational search.

A total of 500 argon atoms were enclosed in a cubic box and the cubic periodic boundary conditions were applied to the system except otherwise noted. The initial crystal structure was created by putting argon atoms at face-centered-cubic (fcc) lattices, and the liquid structure was obtained by heating the crystal to 120 K and equilibrating for 1000 ps. A neighbor list was created for interaction calculations and updated every 10 MD time steps. A time step of 0.01 ps was employed in all our simulations. The cut-off distance for atom-atom interaction was set at 14.0 Å. A switch function was applied between 10 Å and 12 Å to smooth the interaction change across the cutoff.

A system of 500 argon atoms is used in our comparative simulations as shown in Figure 28 tetragonal periodic boundary condition with sides a=b=28.53 Å and c=57.06 Å was applied to the system. Such a boundary condition makes the system look like films of condense phase separated by gas layers. Such a design allows the volume the system to change freely upon crystallization. The starting structure was created by melting a fcc crystal at 120 K, cooling down, and being equilibrated at 60 K. The equilibrated argon liquid film was simulated using the conventional MD method and the SGMD method at t_l=0.2 ps and λ=0.02, 0.05, and 0.1. In the conventional MD simulation, it took 65 ns before crystallization occurred. While in the SGMD simulations, the crystallization occurred at 63, 2, and 0.5 ns with λ=0.02, 0.05, and 0.1, respectively. Figure 29 shows the potential energy changes during these simulations. Phase transitions are evident by the sharp decline in potential energy.

For the crystallization of argon liquid, the SGMD simulations result in a dramatic acceleration in searching out the low free energy states as compared to conventional MD simulations.

Example 5

The present example illustrate employing the method of the invention in determining the binding modes of benzyl alcohol with -cyclodextin in solution. This system models complexation of -cyclodextin with aromatic guests in general and also provides a model for aromatic-alkyl alcohol guests, which have been widely used as antiseptics.

Two Self-guided Molecular Dynamics (SGMD) simulation runs were performed on the -cyclodextin – benzyl alcohol system, simulation I, 1.5 ns and simulation II, 1.0 ns long. Both simulations used the same conditions: the time step was 0.002 ps, the temperature 300 K; non-bonding cutoff was 14 Angstrom and a smoothing function was applied between 8.0 and 12.0 Angstroms. The

dielectric constant was set to one, the frequency of updating the non-bonding pair list was every 20 steps and the self-guiding force was used with 0.1 scaling factor and 0.2 ps averaging time. The system was solvated using periodic boundary conditions (TIP3 water) with box size 35.0 Angstrom in each directions. The Charmm22 force field was used.

In the starting conformation six benzyl alcohol molecules were placed around cyclodextrin. In the case of simulation I, four benzyl alcohol molecules surrounded the rim of the cyclodextrin's macro-ring, two molecules being 'above' and 'below' of the host's cavity. In the case of simulation II all six benzyl alcohol molecules were placed around the rim, approximately in the plane of the glycosidic oxygens. In all cases, the aromatic ring of the benzyl alcohols were oriented approximately parallel to the latter plane.

The conformations were saved every ps in each simulation and then grouped into clusters. In simulation II there are two benzyl alcohol guest molecules bound to cyclodextrin in different regions of the trajectory. Consequently the conformations are considered twice, with respect to the first guest and the second guest case yielding a total of 3500 conformations.

The structures were clustered based on the following criteria: (a) center of mass distance between host and guest considering heavy atoms only; (b) the number of hydrogen bonds between the benzyl alcohol guest and -cyclodextrin's secondary hydroxyls or (c) water molecules; (d) the distance between the center of mass of the oxygens in the secondary hydroxyls and the polar end of the guest. Here the polar end was the center of mass of the oxygen in benzyl alcohol and the carbon atom bonded to; (e) the angle between the plane of the glycosidic oxygens and that of the guest's phenyl ring.

The benzyl alcohol molecules in the starting conformation are distant from the host's cavity as shown in Figure 30. During the SGMD simulations, reversible

guest binding as well as competitive binding of two benzyl alcohol molecules were observed. Since host-guest unbinding and binding occurs in both simulations several times, it was likely that the length of these simulations was sufficient for adequate sampling of the conformational space of the complex.

The conformations collected during simulation I and II are grouped into clusters in order to determine favored and less important binding modes. One of the major binding modes shows high similarity with that observed in the available crystal structure, as discussed below. Also there are other features of benzyl alcohol binding to cyclodextrin characteristic to the solution complex, to which SGMD conformational analysis was able to give detailed insight.

There are two major binding modes identified from the simulation I, II trajectories while all other cluster conformations are much less represented. The two major clusters appear to share common characteristics as follows. The guest molecule does not hydrogen bond with the host's secondary hydroxyls, similarly to the conformation in the crystal structure. In the latter case the guest hydrogen bonds with other cyclodextrin molecules packed close in the crystal in a herring-bone like pattern. Even though conformations of the first cluster contain the guest fully inserted into the host's cavity, there was extensive hydrogen bonding between the complexed guest and the solvent similarly to the other major and overwhelmingly to most minor clusters. Interestingly, a minor third cluster exhibits the lowest non-bonding interaction energy with cyclodextrin most likely due to satisfying fully its hydrogen bonding potential through interactions with the host as well as with the solvent. However, hydrogen bonding between the host and guest does not necessarily lead to lowering of the energy of the complex due to likely loss of hydrogen bonding between the host and water molecules.

The orientation of the benzyl alcohol in the major clusters was similar to that observed in the X-ray structure. The phenyl ring was not perpendicular to the

plane of the glycosidic oxygens; it forms an angle around 70°, which was very close to the value (69.6 °) observed in the crystal structure of the complex.

The most important difference between the two major binding modes (major clusters) was the fact that the first cluster was more deeply inserted into -clyclodextrin's cavity compared to the third cluster and hence it was closer to the degree of inclusion characteristic of the crystal structure. For comparison purposes, we calculated the center of mass distance between host and guest (heavy) atoms for the X-ray structure obtained from the Cambridge Structural Database (CSD). The latter distance was 0.53 Angstroms, comparable to the average value (0.77 ±0.28) of the first major cluster.

Comparing the host's conformation between the two major cluster complex structures obtained in this example, there are no significant differences. In contrast with their overall averages, the average maximum and average minimum of the distances of the glycosidic oxygens from their center differ comparing the non-bonding clusters with the major two clusters, consistent with slight elliptical distortion of the clyclodextrin macro-ring upon complexation.

Minor clusters obtained from the trajectories of simulation I and II contain non-bonding clusters, which represent conformations of clyclodextrin and benzyl alcohol while there was considerable distance between them. In additions to the case of one guest molecule complexed with the host (like the crystal structure), the simulation trajectories indicate competitive and reversible binding between clyclodextrin and benzyl alcohol. The distribution of the conformations among the clusters clearly show that the orientation of the benzyl alcohol with its hydroxyl group toward the host's secondary hydroxyls (toward the wider side) was not favorable (also referred to as the 'incorrect' orientation as opposed to the 'correct' orientation). Clusters having the guest in this orientation are very poorly represented. Furthermore, in simulation I, benzyl alcohol initially binds in the

'incorrect' orientation, but after approximately 65 ps, the guest molecule leaves the host's cavity, re-binds and stays in the 'correct' conformation for the next 425 ps simulation time without interruptions. Simulation II trajectory contains also only brief regions of 'incorrect' orientations even though competitive replacement of a benzyl alcohol in cyclodextrin's cavity by another benzyl alcohol molecule occurs in both, 'correct' and 'incorrect' orientations.

Analysis of the SGMD trajectory suggests that inclusion complex formation was hydrophobically driven, since guest binding appears to effect primarily the exposure of the non-polar rather than the polar part of the host-guest system to the solvent. Hence our data suggest that burying of the aromatic hydrophobic moiety of benzyl alcohol makes guest binding favorable, as opposed to host-guest hydrogen bonding.

As illustrated in this example, the invention provides for the first time a computational approach to determine the binding mode of an aromatic guest with cyclodextrin which does not incorporate *a priori* information regarding the guest's orientation or position with respect to the host. One of the two major binding modes determined agrees well with the crystal structure available for the studied complex as illustrated in Figure 31. The computational results are also informative of features characteristic of the in solution structure of the cyclodextrin complex.

This example also demonstrates the efficiency of the SGMD method used with a widely accepted force field for applications/cases in which rapid exploration of the available conformational space was essential.

While the invention has been described in terms of preferred embodiments, the skilled artisan will appreciate that various modifications, substitutions, omissions and changes may be made without departing from the spirit thereof.

Accordingly, it is intended that the scope of the present invention be limited solely by the scope of the following claims, including equivalents thereof.